International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp.145-156
http://dx.doi.org/10.14257/ijdta.2015.8.2.14

Measurement and Analysis of Burst Topic in Microblog

Guozhong Dong[1], Xin Zou[2*], Wei Wang[1], Yaxue Hu[1], Guowei Shen[1], Korawit Orkphol[1] and Wu Yang[1]

1Information Security Research Center, Harbin Engineering University Harbin, Heilongjiang Province, China 150001
2National Computer Network Emergency Response Technical Team/Coordination Center, Beijing, China 100029
dongguozhong@hrbeu.edu.cn, zouxin@cert.org.cn, w_wei@hrbeu.edu.cn, huyaxue@hrbeu.edu.cn, shenguowei@hrbeu.edu.cn, korawit.orkphol@gmail.com, yangwu@hrbeu.edu.cn

Abstract

Microblog provides the first communication platform for burst event due to the immediacy and interactivity of microblog. In this paper, we research on user-oriented and message-oriented measurements of burst topic in Sina microblog. The measurements and analysis on large-scale Sina microblog data set show that our proposed measurement method can measure the characteristics of user and message propagation in burst topic. The measurement results in this paper can describe the formation and diffusion mechanism of burst topic which will contribute to better research of relevant issues on burst topic and ensure the well-developed of microblog.

Keywords: *Sina microblog, burst topic, user-oriented measurement, message-oriented measurement*

1. Introduction

With fast development of Web 2.0, social network based on Internet has been one of the most representative applications and one of the most important platforms to obtain information, as the major social network platforms. Microblog is a communication and user-generated platform which makes everybody be producer, communicator and commentator. Since microblog is great source for getting access to information much faster than traditional sources of media, more and more organizations and public figures post and spread information through microblog. With the rapid expansion of information and the rapid increasing of users, microblog has gradually evolved from social media to the public platform to express their feelings and opinions, to discuss public affairs, and to conduct public opinion. Because of immediacy and interactivity of microblog, microblog provides the first communication platform for burst event as soon as event occurs. For example, oil explosion accident in Qingdao, microblog is the earliest news source. The reports and discussions for burst event are benefit to crisis response and situational awareness. But because of spreading messages almost instantly and spotty user quality, it will speed up the propagation velocity and have a negative effect on controlling network public opinion when burst topics are spread by malicious users within a short time. All these issues have brought challenges to study on burst topic in microblog including burst topic detection, mining key users of burst topic, control and prediction of topic diffusion. Faced with these challenges, we research on user-oriented and message-oriented measurements of burst topic in microblog from three points of view: user attribute, user behavior and message propagation. The measurements and analysis on large-scale Sina microblog data set show that our

ISSN: 2005-4270 IJDTA

proposed measurement method can measure the characteristics of user and message propagation in burst topic and can describe the formation and diffusion mechanism of burst topic. The measurement results in this paper contribute to better research of relevant issues on burst topic and ensure the well-developed of microblog.

The rest of the paper is organized as follows. We review the related work in Section 2. Section 3 presents the data collection and description. User-oriented and message-oriented measurements of burst topic are reported in Section 4, and conclude this paper in Section 5.

2. Related Work

User behaviors and information propagation patterns can be better understood by measuring and analyzing microblog network. Domestic and foreign researchers design qualitative or quantitative measurements before studying on user influence[1-7], and information propagation[8-12]. Me young *et al.*[13] present an in-depth comparison of three measures of influence: in degree, retweets, and mention. Based on these measures, the dynamics of user influence across topics and time is investigated. Philip *et al.*[14]focus on measuring user's full potential influence inherent in the user connectivity network and propose a modified k-shell decomposition algorithm for computing user influence on Twitter. Li *et al.*[15] propose Topic-level Opinion Influence Model(TOIM) to predict users' future opinions on some specific topics based on the analysis of users' historical messages and social interaction records. Ye *et al.*[10] conduct research on message propagation and social influence by analyzing the propagation patterns of general messages. Liu *et al.*[16] analyze the relationship between user behavior factors together with the diffusion of social influence and propose a learning-based method to measure user influence via predicting users' capability of propagating information. Aditi *et al.*[17] analyze the credibility of information in tweets corresponding to fourteen high impact news events and identify the important content and sourced based features to predict the credibility of information in a tweet. Yan *et al.*[18] prove that the small-world characteristic of microblog social network and the degree distributions of users are power-law. A social network based on human dynamics model is proposed based on their empirical analysis. Fan *et al.*[19] measure the topological characteristics and user behavior patterns in Sina microblog which are helpful for monitoring and controlling the microblog. Liu *et al.*[20] conduct the topic-oriented research on the measurement from many aspects such as features of the content, the network topology and user behavior. The measurement indicators and results can be effectively applied in a topic-generated network. Existing analysis and measurement in microblog mostly focus on the topological characteristics and propagation patterns of single entity, rather than studying on the relation between different measuring objects in burst topic. To the best of our knowledge, together with the recent studies on social network, we are among the first to make static and dynamic measurement of burst topic in Sina microblog.

3. Data Collection and Description

As the largest microblog platform in China, Sina microblog allows third-party developers to create applications for its openness. We selected Sina microblog as observation platform to measure and analyze burst topics. Considering the characteristic of real-time and huge data, we developed web crawler to collect data set of burst topics. 16 volunteers were interested in our work and participated directly in the data collection process. Before crawling data, volunteers selected 28 burst topics in Sina microblog and labeled hashtags for each burst topic according to the keywords of burst topics. These labeled hashtags were used as keywords to collect messages and user information of burst topic. Messages in spreading process and user information were also crawled with a depth-first strategy. The collected data set covered the period

from August 15 to August 30 in 2014. In order to conduct better analysis and measurement, we selected the most representative burst topic (Jaycee Chan taking drug) as the measurement object. Through data extraction of burst topics, data cleaning and data integration, the data set of selected burst topic contains nearly 600,000 users and over 8 million messages. The representative attributes of collected data set can be seen in Table 1.

Table 1. The Representative Attributes of Collected dataset

Attribute	Attribute description
topic_id	burst topic flag
message_id	message flag
user_id	user flag who posted message
parent_mid	parent message flag in repost network
message_time	post time of message
repost_number	the number of reposts
comment_number	the number of comments
text	the text content of message
username	the username of user
follower_number	the number of user's followers
following_number	the number of user's followings
authenticated	authenticated user flag

4. Measurements and Analysis of Burst Topic

In Section 3, we introduce our data collection strategies and representative attributes of collected data set. This section will describe user-oriented and message-oriented measurements of burst topics in microblog. Also, Figure 1 shows different characteristics and internal relationship among characteristics in burst topic.

4.1. User-Oriented Measurement and Analysis of Burst Topic

Users in microblog play an important role in the evolution of burst topic. We measure and analyze users in burst topic from three aspects: user attribute, user behavior and user influence in this section.

4.1.1. User Attributes: The measurement of user attribute that we focus on can be classified into two categories: the number of followers and user authentication type.

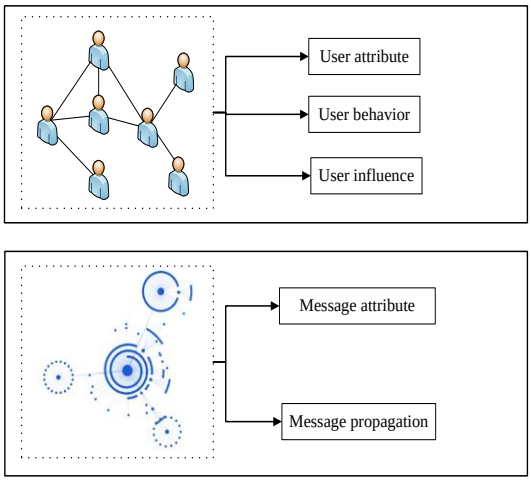


Figure 1. Overview of Measurements of Burst Topic

(1) The Number of Followers

The number of user's followers can reflect user influence to some extent. The distribution of user's followers is shown in Figure 2. As shown in Figure 2, about 85% of users have less than 3,000 followers and 56% of users have 300-3,000 followers. Only a small percentage of users have a large number of followers. These famous users are usually the initiators and guides.

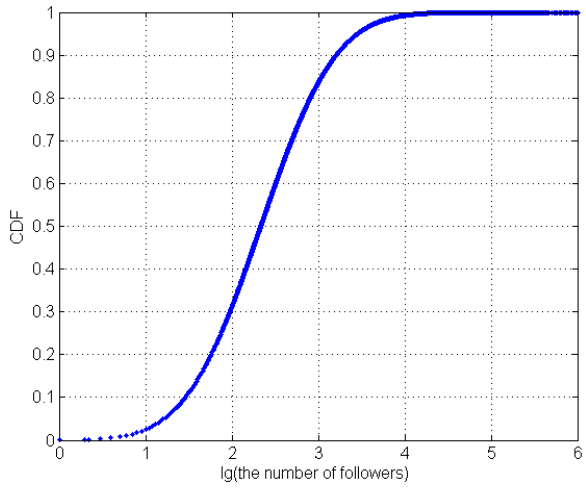


Figure 2. The Distribution of Users' Followers

In order to describe the role of the number of followers in topic propagation, we measure the relation between the number of followers and message popularity where message popularity is defined as the number of reposts and comments. The relation between the number of followers and message popularity is shown in Figure 3. Also, correlation coefficient is used to measure the correlation which can be defined as follows:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

Where r is correlation coefficient, X and Y are the corresponding measured variables. From analyses on the relation between the number of followers and message popularity, their related coefficient is 0. 3993, indicating a weak positive relation. Users who have a large number of followers don't mean that their messages must be reposted or commented widely.

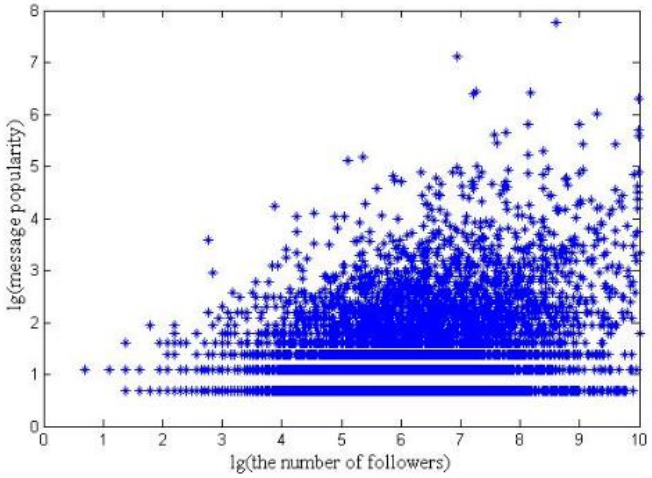


Figure 3. The Relation between Number of Followers and Message Popularity

(2) User Authentication Type

Users in Sina microblog can be divided into two categories: authenticated users and non-authenticated users according to user authentication type. Authenticated users mean that they have higher authenticity and credibility than non-authenticated users.

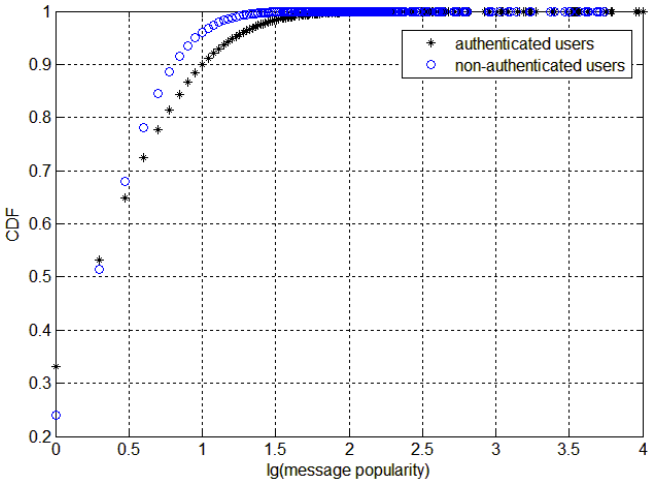


Figure 4. The Distribution of Message Popularity of Different User Authentication Types

We use CDF (Cumulative Probability Distribution) to describe the differentiation of different user authentication types. The distributions of two kinds of user types' message popularity are shown in Figure 4. As shown in Figure 4, only a small percentage of users have high message popularity and almost all messages' message popularity are below than 10. The result shows that the distributions of two kinds of user types' message

popularity are similar, which can't be used to predict message popularity in burst topics based on user authentication type alone.

4.1.2. User Behavior: We focus on two kinds of user behaviors in burst topics in this section: post behavior and interaction behavior. When burst event occurs, users can post original messages or interact with users that posted original messages. In order to measure the time distribution of post behavior which reflects massive users' behavior characteristic in burst topic, we set time window to be 1 hour and the time range to be 72 hours from the day that burst event happened. The time distribution of message in burst topic is shown in Figure 5. As shown in Figure 5, the post behavior not only conforms to bedtime and rising time, but also the evolution of burst topics. There are few messages about burst topics between 1 a.m and 7 a.m. Through the analysis of message content, we can find that users' post behaviors change with the development and evolution of burst topics at other times.

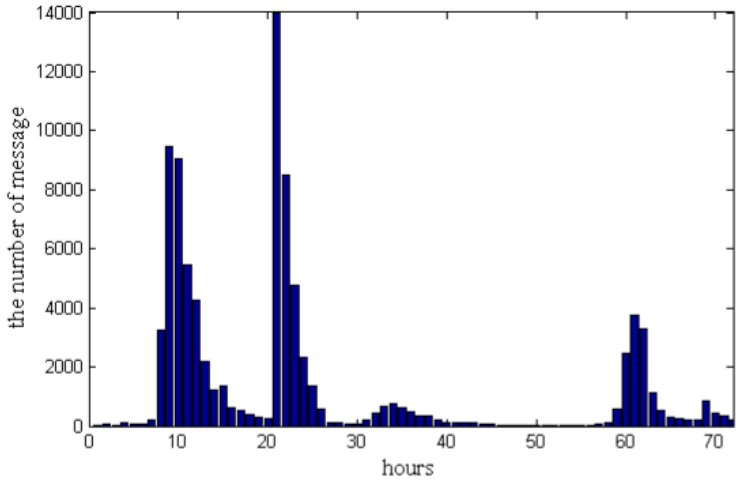


Figure 5. The Time Distribution of Message in Burst Topic

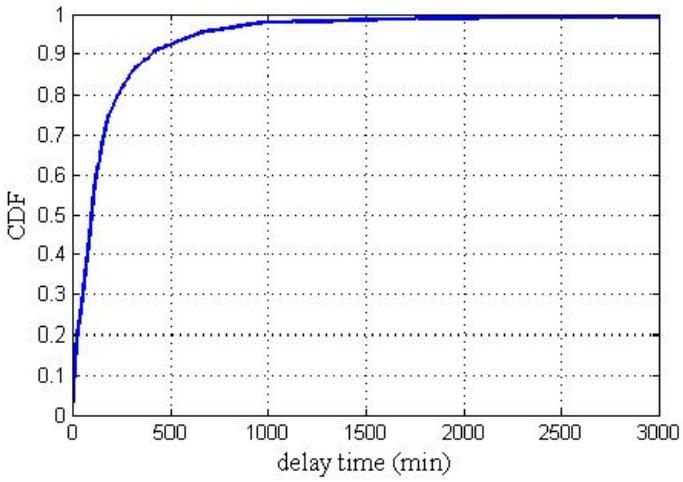


Figure 6. The Distribution of Delay Time in Burst Topic

Besides, we measured the distribution of user interaction delay time in burst topics which is defined as time difference between message and its parent message in message forwarding tree. The distribution of delay time in burst topic is shown in Figure 6.

According to some statistics, 50% of messages' interaction delay time are below 100 minutes and 90% are about 400 minutes. Information propagation in burst topics has stronger timely effectiveness and users don't usually repost and comment outdated messages in burst topic.

4.1.3. User Influence: Influential users can deliver contents to a larger audience than a normal user and have a key role in topic propagation process. User graph is modeled based on the information propagation in burst topic. Nodes in user graph represent users in burst topic. Edges represent repost relation between nodes and the arrows go opposite to the information flow. We leverage link topological ranking by means of PageRank algorithms to detect influential users in burst topics. The PageRank of the node is the sum of contributions from its incoming edges. Damping factor was set to be 0.85 in the measurement of user influence. The ranking result of influential users based on PageRank can be seen in Table 2.

Table 2. The Ranking Result of Influential Users in Burst Topic (Jaycee Chan taking drug)

Ranking	Username
1	Sina Entertainment Video
2	CCTV News
3	Sina Entertainment
4	People's Daily
5	Phoenix Entertainment
6	Headline News
7	Professor Silver

As Table 2 shows, seven of the most influential users mainly are media organizations and famous users in burst topic of Jaycee Chan taking drug. Influential users as an important part of microblog, they play a very important role in promoting the development process of public opinion of burst event.

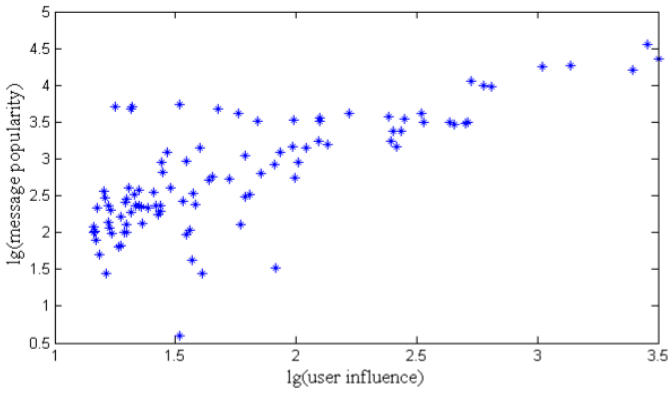


Figure 7. The Relation Between User Influence and Message Popularity

The correlation between user influence and message popularity was measured. The relation between user influence and message popularity is shown in Figure 7. Related coefficient between user influence order and message popularity is 0.9002, indicating a

strong positive relation. Detecting influential users contribute to research on predicting topic trend and topic tracking.

4.2. Message-Oriented Measurement and Analysis of Burst Topic

A large scale message-oriented measurements and analysis in burst topic are performed to better grasp the characteristics of message and message propagation pattern in this section.

4.2.1. Message Attributes: In the process of topic propagation, some messages have a significant increase during a certain time interval which is the major cause of burst topic. Six representative spikes of online media are shown in [21]. We can detect burst messages based on message's rise-and-fall patterns. To model message's burst attribute, messages in message forwarding tree will be divided into different time windows based on the post time of message to obtain the time sequence of message. The burst detection approach based on technical analysis indicators such as EMA, MACD, and MACD histogram is defined in [22]. A burst is defined as a time interval in which the MACD Histogram value is greater than burst threshold. The strength of the burst at a specific timestamp is the MACD Histogram value of that timestamp. MACD Histogram values of different time windows are shown in Figure 8. The message has burst attribute between the eighth and twelfth time window when burst threshold is set to be 30. Through analysis of users that posted burst messages, we found that burst messages are the key nodes in the development of burst topic. Detecting burst message in real-time can contribute to detect burst topic and key users in burst topic.

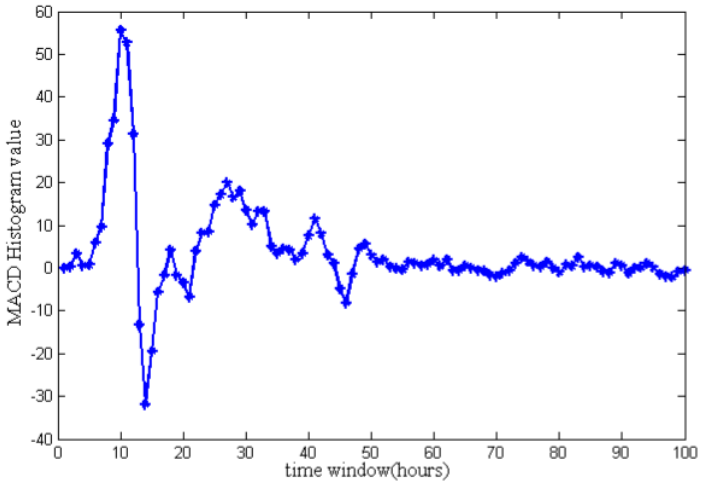


Figure 8. MACD Histogram Values of Different Time Windows

4.2.2 Message Propagation: Through the analysis of influential users' message forwarding tree as shown in Figure 9, the propagation topology of forwarding tree can be divided into two categories: star propagation and tree propagation. The sources of star propagation are almost the users that post original messages and trendsetters in burst topic. Most of users in star propagation structure are source's followers and mainly in the first level of propagation tree. Tree propagation structure may consist of more than one influential user. Each propagation structure of influential user can be seen as star propagation. Based on the analysis of the propagation topology, the width and depth of message propagation were measured. We found that the depth of forwarding tree is small and almost all of messages' depths are below 6. However, the widths of popular messages are large, especially compared with the depths.

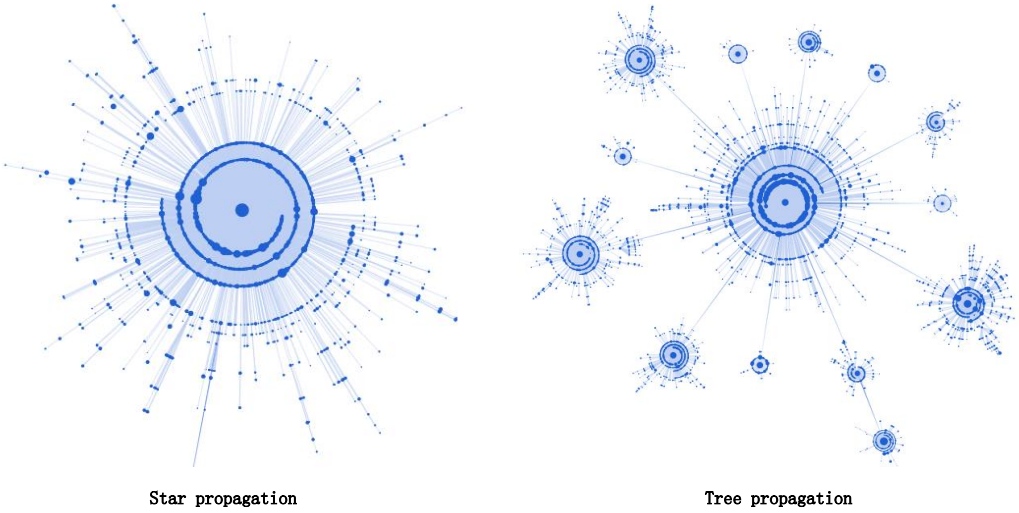


Figure 9. The Propagation Topology of Forwarding Tree

5. Conclusion

In this paper, we have addressed the problem of measurement and analysis of burst topic in microblog. An effective measurement approach based on user entity and message entity is presented. We measure and analyze burst topic from five aspects: user attribute, user behavior, user influence, message attributes and message propagation. Besides, the correlation analysis between different aspects are measured. To the best of our knowledge, together with the recent studies on social network, we are among the first to make static and dynamic measurement of burst topic in Sina microblog. The measurements and analysis on large-scale Sina microblog data set show that our proposed measurement method can contribute to better research of relevant issues on burst topic.

Acknowledgements

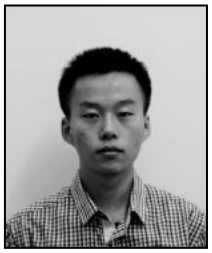
The authors would like to thank the reviewers for suggesting many ways to improve the paper. This work was partially supported by the National High Technology Research and Development Program of China(no. 2012AA012802), the National Natural Science Foundation of China(no. 61170242, no. 61101140, no. 61272537) and the Fundamental Research Funds for the Central Universities(no. HEUCF100611).

References

[1] J. Li, P. W. Li, T. Sun, T. Li and Q. X. Jian, "Social network user influence sense-making and dynamics prediction", Expert Systems with Applications, vol. 41, no. 11, **(2014)**.
[2] Q. Fang, S. Jitao, X. Changsheng and R. Yong, "Topic-sensitive Influencer Mining in Interest-Based Social Media Networks via Hypergraph Learning", IEEE Transactions on Multimedia, vol. 16, no. 3, **(2014)**.
[3] B. Bi, T. Yuanyuan, S. Yannis, B. Andrey and C. Junghoo, "Scalable topic-specific influence analysis on microblogs", Proceedings of the 7th ACM international conference on Web search and data mining Pages, **(2014)** February 24-28, New York, USA.
[4] C. Xiao, Z. Yuhong, Z. Xue and Wu, Yue, "Predicting User Influence in Social Media", Journal of Networks, vol. 8, no. 11, **(2013)**.
[5] S. Singh, M. Nishchol and S. Sanjeev, "Survey of various techniques for determining influential users in social networks", 2013 International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), **(2013)** March 25-26, Tirunelveli, India.
[6] X. Li, C. Shaoyin, C. Wenlong and J. Fan, "Novel user influence measurement based on user interaction in microblog", Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, **(2013)** August 25-28, Niagara Falls, Canada.

[7] N. Barbieri, B. Francesco and M. Giuseppe, "Topic-aware social influence propagation models", Knowledge and Information Systems, vol. 37, no. 3, **(2013)**.
[8] S. Lin, W. Fengjiao, H. Qingbo and Y. S. Philip, "Extracting social events for learning better information diffusion models", Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2013)** August 11-14, Chicago, USA.
[9] L. Weng, R. Jacob, P. Nicola, G. Bruno, A. Castillo, C. Bonchi, F. Schifanella, R. Menczer and F. F. Alessandro, "The role of information diffusion in the evolution of social networks", Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2013)** August 11-14, Chicago, USA.
[10] S. Ye and W. Felix, "Measuring message propagation and social influence on Twitter. com", Int. J. Communication Networks and Distributed Systems, vol. 11, no. 1, **(2013)**.
[11] P. Fan, L. Pei, J. Zhihong, L. Wei and W. Hui, "Measurement and analysis of topology and information propagation on sina-microblog", The role of information diffusion in the evolution of social networks. 2011 IEEE International Conference on Intelligence and Security Informatics, **(2011)** July 10-12, Beijing, China.
[12] K. Saito, K. Masahiro, O. Kouzou and M. Hiroshi, "Selecting information diffusion models over social networks for behavioral analysis", Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases, **(2010)** September 20-24, Catalonia, Spain.
[13] M. Cha, H. Hamed, B., Fabricio and G. P Krishna, "Measuring User Influence in Twitter: The Million Follower Fallacy", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, **(2010)** May 23–26, Washington, USA.
[14] E. Phil and B. F. Junlan, "Measuring user influence on twitter using modified k-shell decomposition", Fifth International AAAI Conference on Weblogs and Social Media, **(2011)** July 17–21, Barcelona, Spain.
[15] D. Li, S. Xin, S. Guozheng, T. Jie, D. Ying and L. Zhipeng, "Mining topic-level opinion influence in microblog", Proceedings of the 21st ACM international conference on Information and knowledge management, **(2012)** October 29–November 2, Hawaii, USA.
[16] Y. Liu, M. Zhang, S. Ma and J. Mao, "Social Influence Analysis for Microblog User Based on User Behavior", vol. 37, no. 4, **(2014)**.
[17] A. Gupta and K. Ponnurangam, "Credibility ranking of tweets during high impact events", Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, **(2012)** April 17, Lyon, France.
[18] Q. Yan, W. Lianren and Z. Lan, "Social network based microblog user behavior analysis", Physica A: Statistical Mechanics and its Applications, vol. 392, no. 7, **(2013)**.
[19] P. Fan, H. Wang, Z. Jiang and P. Li, "Measurement of Microblogging Network. Journal of Computer Research and Development", vol. 49, no. 4, **(2012)**.
[20] W. Liu, L. Wang and R. Li, "Topic-oriented measurement of microblogging network. Journal on Communications", vol. 34, no. 11, **(2013)**.
[21] Y. Matsubara, S. Y. Prakash, B. Aditya, L. Lei and F. Christos, "Rise and fall patterns of information diffusion: model and implications", Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2012)** August 12-16, Beijing, China.
[22] D. He and P. D. Stott, "Topic dynamics: an alternative model of bursts in streams of topics", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2010)** July 25-28, Washington, USA.

Authors



Guozhong Dong, born in 1989. PhD candidate at Harbin Engineering University. His main research interests include data mining, social computing, etc. (dongguozhong@hrbeu.edu.cn)

Xin Zou, He received master degree from Department of Science and Technology from Harbin Institute of Technology, China 2003. Since 2008, he has worked in National Computer Network Emergency Response Technical Team/Coordination Center of China. His research interests focus on network security, networking optimization and big data mining. (zouxin@cert.org.cn)

Wei Wang, born in 1974. PhD and associate professor at Harbin Engineering University. His main research interests include data mining, information security, etc. (w_wei@hrbeu.edu.cn)

Yaxue Hu, born in 1991. Master Degree Candidate at Harbin Engineering University. Her main research interests include data mining, social network analysis, etc. (huyaxue@hrbeu.edu.cn)

Guowei Shen, is currently a Ph.D candidate in the Department of Computer Science and Technology, Harbin Engineering University. He received his B.E. degree in 2009 from the Department of Computer Science and Technology of Harbin Engineering University, Harbin, China. His main research interests include data mining, social computing and information security. (shenguowei@hrbeu.edu.cn)

Korawit Orkphol, born in 1987. He is a lecturer at Computer Engineering Department, Kasetsart University, Si Racha Campus, Thailand. Now he is a PhD Candidate at Harbin Engineering University. His research interests focus on web technologies, network security, and social network analysis. (korawit.orkphol@gmail.com)

Wu Yang, born in 1974. Professor and PhD supervisor at Harbin Engineering University. His main research interests include data mining, information security, etc. (yangwu@hrbeu.edu.cn)

